Elsevier LaTeX template*

Elsevier[1]

Radarweg 29, Amsterdam

Elsevier Inc[a,b], Global Customer Service[b,]*

[a] *1600 John F Kennedy Boulevard, Philadelphia*
[b] *360 Park Avenue South, New York*

Abstract

This template helps you to create a properly formatted LaTeX manuscript.

Keywords: `elsarticle.cls`, LaTeX, Elsevier, template

2010 MSC: 00-01, 99-00

1. Introduction

Since the dawn of internet and world wide web, humanity has witnessed a degree of connection beyond reckoning. The proliferation of digital devices pervaded with various applications that account for almost all aspect of humanity, have created cyber communities that constantly mutate [1]; [2]. In a world where we have network infrastructures that can support up to 250Mbps of data transmission, and smart phones and IOT devices that can have processing power of up to 3 Ghz, data becomes ubiquitous, the quantum that lays the foundation of the nexus [3].

According to InternetLiveStates.com [4], only in one second, there are 9,878 tweets sent, 1,138 instagram photos uploaded, 3,117,720 emails sent, 99,738 Google searches made, and 94,144 Youtube videos viewed. That is, if it has

*Fully documented templates are available in the elsarticle package on CTAN.

*Corresponding author
Email address: `support@elsevier.com` (Global Customer Service)
URL: `www.elsevier.com` (Elsevier Inc)

[1] Since 1880.

taken 5 second the read the preceding paragraph, during that time, 15,588,600 emails are sent.

Driven by the ambition to harness the power of this deluge of data, the term 'Big Data' (BD) was coined [5]. BD initially emerged to address the challenges associated with various characteristics of data such as velocity, variety, volume and variability [2]. BD is the practice of extracting patterns, theories, and predictions from a large set of structured, semi-structured, and unstructured data for the purposes of business competitive advantage [6]; [7]. BD is a game-changing innovation, heralding the dawn of a new data-oriented industry.

Nonetheless, BD is not a magical wand that can enchant any business process. While a lot of opportunities exist in BD, subsuming an emergent and rather high-impacting technology like BD to current state of affairs in organizations, is a daunting task. According to recent survey from Databricks, only 13% of the organizations excel at delivering on their data strategy [8]. Another survey by NewVantage Partners indicated that only 24% organization have successfully gone data-driven [9]. This survey also states that only 30% of organizations have a well established strategy for their big data endeavour. In addition, surveys from McKinsey & Company ([10]) and Gartner ([11]) further support these numbers, which illuminates on the scarcity of successful big data implementations in the industry.

Among the challenges of data adoption perhaps the most highlighted are 'data engineering complexities', 'big data architecture', 'rapid technology change', 'lack of sufficient skilled data engineers', and 'organization's cultural challenges of becoming data-driven' [2];[12]. This focus of this study is on data engineering complexities and in specific big data architecture.

In the past, organization relied on a few technology giants to provide infrastructure and tools necessary for big data, while today there's a plethora of choice from hundreds of providers covering different aspect of data ecosystem from ingestion, to logging, to stream processing, and to visualization [9]. Companies are tending more and more towards Cloud-native architectures for cost reduction, improved efficiency and new roles have been introduced such as chief

analytics officer (CAOs) amd chief data officers (CDOs) to channel the organizational big data capabilities toward business value and competitive advantage [13].

So how can one embark on this rather sophisticated journey? what can be a good logical approach to absorb the ever-increasing complexity of big data systems? how can organizations build different stacks to handle data for various workloads such as machine learning (ML), business analytics, data engineering, and streaming?

We suggest that majority of the challenge discussed starts with data architecture [1]; [3]. The data ingestion, processing and consumption of different data workloads vary, and sometimes they don't go well together. A company that enacted a data lake and a data warehouse and tries to account for both ecosystems, can be dealing with immense complexity, which in turns impact data teams, which in turn can hinder innovation, create barriers and result in monumental lost.

Development and deployment of an efficacious big data system is only the beginning of a big data journey. As data sources increase, variety of data increases, number of data consumers increase, the data store gets confuscated, and this can introduce threats for scalability and maintainability of the system. This also implies that only a handful of hyper-specialized data engineers would understand the system internals, creating silos, and potential miscommunication.

Majority of these systems are developed on-premise as ad-hoc complicated solutions that do not adhere to the practices of software engineering and software architecture [14]; [15]. As the ecosystem grows and new technologies and data processing techniques are introduced, the software architect will have a harder time to come up with a solution that address the problem requirements.

This can potentially create grounds for an immature architecture that results in solutions that are hard to scale, hard to maintain, and raise high-entry blockades [3]. Since the approach of ad-hoc design to big data system development is not desirable and may leave many architects and data engineers in the dark,

novel data architectures that are designed specifically for BD are required. To contribute to this goal, we explore the notion of reference architectures (RAs) and present a distributed domain-driven software RA for big data systems.

2. Why reference architecture?

To justify why we have chosen reference architectures as the suitable artefact, first we have to clarify two assumptions;

1. having a sound software architecture is essential to the successful development and maintenance of software systems
2. there exist a sufficient body of knowledge in the field of software architecture to support the development of an effective RA

One of the focal tenets of software architecture is that every system is developed to satisfy a business objective, and that the architecture of the system is a bridge between abstract business goals to concrete final solutions [16]. While the journey of big data can be quite challenging, the good news is that a software RA can be designed, analyzed and documented incorporating best practices, known techniques, and patterns that will support the achievement of the business goals. In this way, the complexity can be absorbed, and made tractable.

Practitioners of complex systems, software engineers, and system designers have been frequently using reference architectures to have a collective understanding of system components, functionalities, data-flows and patterns which shape the overall qualities of system and help further adjust it to the business objectives [17]; [18]. There is a fair amount of literature on reference architectures, and whereas different authors definition may vary, they all share the same tenets.

A reference architecture is amalgamation of architectural patterns, standards, software engineering techniques that bridge the problem domain to a class of solutions. This artefact can be partially or completely instantiated and prototyped in a particular business context together with other supporting arte-

fact to enable its use. RAs are often created from previous RAs and architecture [1].

The usage of RAs for the development of complex systems is not new. In software product line (SPL) development, RAs are generic artifacts that are configured and instantiated for a particular domain of systems [19]. In software engineering, major IT giants like IBM has referred to RAs as the 'best of best practices' to address unique and complex system development challenges [17].

Based on the premises discussed and taking all into consideration, RAs can facilitate the issues of big data architecture and data engineering because of the following reasons;

1. RAs can promote adherence to best practice, standards, specifications and patterns
2. RAs can endow the data architecture team with openness and increase operability, incorporating architectural patterns that ensue desirable predefined quality attributes
3. RAs can be the best initial start to the big data journey, capturing design issues when they are still cheap
4. RAs can bring different stakeholders on the same table and help achieve consensus around major technological constructs
5. RAs can be effective in identifying and addressing cross-cutting concerns
6. RAs can serve as the organizational memory around design decisions, enlightening next subsequent decisions
7. RAs can act as a summary and blueprint in the portfolio of software engineers and architect, resulting in better dissemination of knowledge

3. Research Methodology

There are a few studies that have addressed the systematic development of reference architectures. Cloutier et al [17] present a high-level model for RA development through collection of contemporary information and capturing the essence of architectural advancements. In another effort,PuLSE-DSSA

is proposed by Bayer et al. [20] in the context of product line development and domain engineering. PulSE-DSSA emphasizes on capturing the existing architectural knowledge. Stricker et al. [21] propose a pattern-based approach for creating an RA. This study revolves around software engineering patterns motivated by the work of Gamma et al [22]; proposing a structural approach that includes three layers of patterns with well-defined hierarchical relationships. Nakagawa, Martins, Felizardo, and Maldonado [23] propose an approach to RA design outside of product line management context that is concentrated towards aspect-oriented systems.

Galster and Avgeriou [24] propose an empirically grounded reference architecture based on two main facets; Existing RAs in practice and available literature on RAs. Along the same vein, Nakagawa et al [25] presented ProSA-RA which is a 4 phase methodology that unlike many other methodologies do provide a more comprehensive instructions on RA evaluation. In addition, this methodology benefits from an ecosystem of complementary constructs that aid in RA design and evaluation such as RAModel [26] and a framework for evaluation of RAs (FERA) [27].In a recent study, Derras et al. [28] propose a schema of practical RA development in the context of software product line and domain engineering. This study is based on capturing knowledge from architectures in practice with attention to variability, configurability and product line development. The findings provide a four-phase process to develop quality driven reference architectures. This approach is influenced by ISO/IEC 26550 [29].

By analysis and study of all these approaches for design and development of RAs, a common pattern has been witnessed. Whereas some of them are more recent and some belong to years ago, there are commonalities that has been observed. All these approaches are grounded on three main pillars, 1) Existing RAs 2) RAs in literature 3) Architectures in practice. Taking this into consideration and by analyzing the results of the systematic literature review conducted by Ataei et al [1] we found'Empirically-grounded reference architectures' proposed by Galster and Avgeriou [24], a suitable methodology, because firstly it's been adopted by many studies, and secondly it's comparatively in-line with the

nature of our study.

Nevertheless, we did not fully adopt this methodology and rather customized to the needs of this particular research. This is due to some inherent limitations that has been witnessed with the methodology. For instance we could not find a comprehensive guideline on how to identify data sources and how it could be categorized and synthesized into the creation of the RA in the third step of the methodology, therefore we employed the Nakagawa's information source investigation guidelines and the overall idea of the RAModel. Another limitation we've faced was with evaluation of the RA. As evaluation, second to a sound research methodology is one of the key elements of any good design science research, we had to look for a stronger and more systematic evaluation approach than what was discussed in 'empirically grounded RAs' methodology. For this purpose, and inspired by the works of Angelov et al [30]; [31], we first created an prototype of the RA in practice, and then used 'The architecture tradeoff analysis method' (ATAM) [32] to evaluate the artefact.

This research methodology is constituent of 6 phases which are respectively; 1) Decision on the type of the RA 2) Design strategy 3) Empirical acquisition of data 4) Construction of the RA 5) Enable RA with variability 6) Evaluation of the RA. The phrase 'empirically grounded' refers to two major elements; firstly the reference architecture should be grounded in well-established and proven principles; secondly, the reference architecture should be evaluated for applicability and validity. These don't only belong to Galster and Avgeriou methodology, and other researchers such as Cloutier [17] and Derras et al [19] have promoted the same ideas.

It is worth mentioning that this methodology is iterative, meaning that the results gained from the evaluation phase (6th phase) determines the subsequent iterations until the design reaches saturation.

3.1. Step1: Decision on type of the RA

Precursor to any effective RA development, is the decision on type of it. The type of the RA is significant, as it illuminates on information to be collected

and the construction of the RA in later phases. The selection on the type of RA for the purposes of this study is based on two dimensions; the classification framework proposed by Angelov et al. [33] and the usage context [34].

Based on the classification framework proposed by Angelov et al. [33], five types of RA are defined. This framework has been developed with the goal of supporting analysis of RAs with regards to context, goal, and the architecture specification/design relationships. It is based on 3 major dimensions namely context, goals, and design, each having their own corresponding sub-dimensions. These dimensions and sub-dimensions are derived by the means of interrogatives (the usage of interrogates is a well-established practice for problem analysis (the usage of interrogates is a well-established practice for problem analysis).

The interrogatives 'When', 'Where', and 'Who' have been used to address the 'context', 'Why' has been used to address 'goal', and 'How' and 'What' have been used to address 'design' dimension. The outcome of the study categorizes RAs in two major groups; 1) standardization RAs and 2) Facilitation RAs. This framework has been chosen because it is completely in-line with the purposes of this study and aims to demarcate a clear domain for the RA to be developed. The comprehensive classification of the RAs with examples in practice illuminates on how different RAs are playing roles in the industry and how they are classified. This brings clarity on what should be developed and what boundaries should be drawn.

By reading the results of the recent SLR conducted by Ataei et al on BD RAs [1], we've added more examples of the RAs on top of what was provided by Angelov [33], and provided the following updated list of RA classifications with examples;

1. Standardization RAs
   (a) Type 1: classical, standardization architectures designed to be implemented in multiple organizations. Examples are:
      i. WRM [35]
      ii. OSI RM [36]

iii. OATH [37]
iv. COBRA [38]
v. Neomycelia [3]
vi. Kappa [39]
vii. Bolster [15]

(b) Type 2: classical, standardization architectures designed to be implemented in a single organization
i. Fortis Bank Reference Software Architecture [**?**]

2. Facilitation RAs

(a) Type 3: classical, facilitation reference architectures for multiple organizations designed by a software organization in cooperation with user organizations
i. Microsoft Application Architecture for .Net [40]
ii. IBM PanDOORA
iii. OATH [37]
iv. COBRA [38]

(b) Type 4: classical, facilitation architectures designed to be implemented in a single organization
i. Achmea Software Reference Architecture [41]
ii. ABN-AMRO Web Application Architecture [42]

(c) Type 5: preliminary, facilitation architectures designed to be implemented in multiple organizations
i. ERA [34]
ii. AHA [43]
iii. eSRA [44]

The domain driven distributed BD RA chosen for the purposes of this study pursues two major goals; 1) enabling and support the development and data engineering of big data systems 2) concurrently ensuring that interoperability between different heterogeneous components of the big data system is established. Therefore, the outcome artefact will be a BD RA that is a classical standardization RA designed to be implemented in multiple organizations.

3.2. Step2: Selection of Design Strategy

Angelov et al [30] and Galster et al[24] have both presented that RAs can have two major design strategies to them; 1) RAs that are designed from scratch (practice driven), 2) RAs that are based on other RAs (research driven). Designing RAs from scratch is rare, and usually takes place in an emergent domain that have not perceived a lot of attention. On the other hand, most RAs today are the amalgamation of a priori concrete architectures, models, patterns, best practices, and RAs, that together provide a compelling artefact for a class of problems.

RAs developed from scratch tend to create more prescriptive theories, whereas RAs developed based on available body of knowledge tends to provide with more descriptive design theories. The RA designed for the purposes of this study is a research-based RA based on existing RAs, concrete architectures, and best practices.

3.3. Step 3: Empirical Acquisition of Data

As aforementioned, due to the limitation witnessed by this research methodology, we have augmented this phase, and increase the systematicity and transparency of data collection and synthesis through various academic methods such as systematic literature review or SLR.

This phase is made up of three major undertakings; 1) identification of data sources; 2) capturing data sources; 3) synthesis of data sources.

3.3.1. Identification of data sources

To identify suitable data sources, we've employed the first step of ProSA-RA methodology, 'information source investigation'. This step is an endeavour to capture focal and ancillary knowledge and theories that revolve around the target domain, and lay the ground of RA development.

To unearth the architectural quanta, and to highlight gradations between various approaches to BD system development, we've selected most relevant sources as the followings;

1. **Practice-led conferences:** given that majority of recent advancements for emerging technologies such as microservices architecture [45];[46] [47] and big data are coming from virtually hosted practice-led conferences, we've chosen some of the best conferences hold world-wide for the purposes of data collection. These conferences are 1) Qcon [48] 2) State of Data Mesh by ThoughtWorks [49] 3) Worldwide Software Architecture Summit'21 [50] and 4) Kafka Summit Europe 2021 [51]. Our objective was to capture the frontiers of software architecture and emerging approaches currently being practiced in IT giants such as Google, Facebook and Netflix. Among all the speech in these conferences, we looked for topics that entailed the keywords 'emergent software architecture trends', 'distributed software architecture', and 'big data software architecture'. We used the software Nvivo to code the transcripts from the conference videos. We used the aforementioned keywords as the codes and associated different texts, summative, essence-capturing sentences, evocative attributes to them. During this process, a new theme 'domain driven design' emerged. We added that into the list of codes as well.
2. **Publications:** in order to capture evidence from the body of knowledge, we conducted a systematic literature review (SLR), following the guidelines of PRISMA presented by Moher et all [52]. The main objective of this SLR was to highlight common architectural constructs found among all the BD RAs. This SLR is build on top of our recent work [1] that covered all the RAs by 2020.

   The initial SLR included IEEE Explore, ScienceDirect, SpringerLink, ACM library, MIS Quarterly, Elsevier, AISel as well as citation databases such as Scopus, Web of Science, Google Scholar, and Research Gate. The SLR search keywords used were 'Big Data Reference Architectures', 'Reference Architectures in the domain of Big Data', and 'Reference Architectures and Big Data'. We followed the exact methodology, but this time for the years 2021 and 2022. Our aim was to find out if there has been any new BD RA published during the years mentioned.

By the result of this SLR, we've found 3 more BD RAs ([3]; [53]; [54]) and we've added two new standards ([55]; [56]) to further solidify our study. Converging these new SLR with the old, covering the years 2010-2022, we've pooled 89 literature in the primary phase, and another 10 by snowballing and citation searching. These 99 literature then went through our inclusion, exclusion and quality criteria. These criteria is as blow;

- Inclusion criteria:
  (a) studies that entailed real-world scenarios or tend to solve a problem in practice
  (b) qualitative or quantitative researches that intended to solve the industry gaps in big data system development and architecture
  (c) studies that had strong evaluations, preferably those that created the artefact in an actual organizational setup
  (d) explores the concept of RAs
  (e) provides or builds up on thorough discussion on BD RAs, limitations, drivers, and the overall ecosystem
  (f) is recent, within the years specified
  (g) is a conference paper, journal paper, book, book chapter, white paper, dissertation or thesis
- Exclusion criteria:
  (a) the study is not well evaluated or practice driven
  (b) is a duplicate
  (c) the study is not in-line with research objectives
  (d) not written in English
  (e) provides a poor quality or misleading technical information
- EQuality assessment:
  (a) is the study rich in terms of relevance to practice?
  (b) does the study create related design/design science or kernel theories?
  (c) does the study entail sufficient data?

(d) does the study discuss the recent trends in BD domain?

(e) is the study based on primary data and is internationally focused?

3.3.2. Data Synthesis

After pooling the studies, we removed 11 studies before screening because either they were duplicates or they were not in English. The remaining 78 studies went through screen, in which, 2 studies excluded based on the exclusion criteria. From there on, 76 studies have been assessed for eligibility based on the quality framework and the inclusion criteria. The result of this process handed over 67 studies from this branch. From the other branch, 10 records identified through citation searching. These reports have been assessed through the same quality framework, inclusion and exclusion criteria, which yielded 5 studies from this stream. Together 68 studies pooled for this SLR as depicted in figure 1.

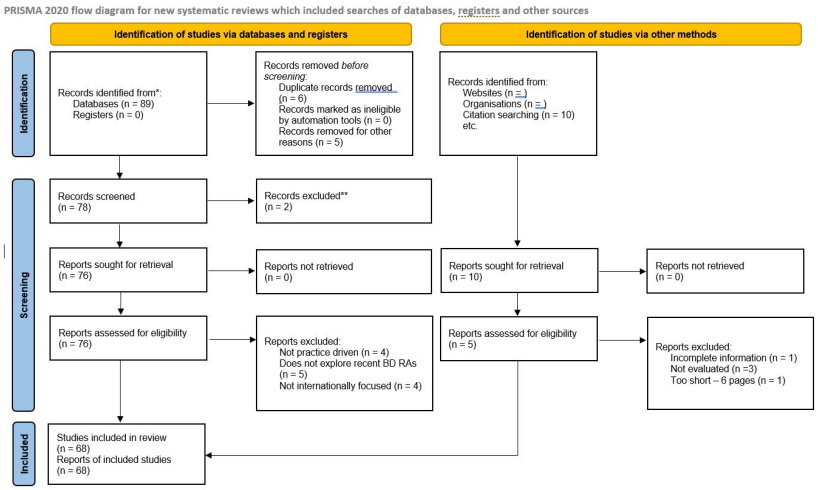

Figure 1: PRISMA flowchart

These 68 studies are comprising of journal papers, conference papers, book chapters, tech reports, tech surveys white papers, standards, master thesis, and PhD dissertations. Out of the pool of these studies, 39.4% are from IEEE Explore, 4.4% are from ScienceDirect, 23.5% are from Springerlink, 13.2% are from

ACM, and 29.4% are from other sources such as citation search, Google Scholar and Research Gate. 30 journal articles, 14 conference papers, 6 whitepapers, 2 ISO standards, 14 book chapters, and 2 postgraduate studies have been selected. 26% of these studies are from the year 2010-2013, 33% are from the years 2013-2015, and 51% are from the years 2016-2022. These stats are portrayed in figure 2.

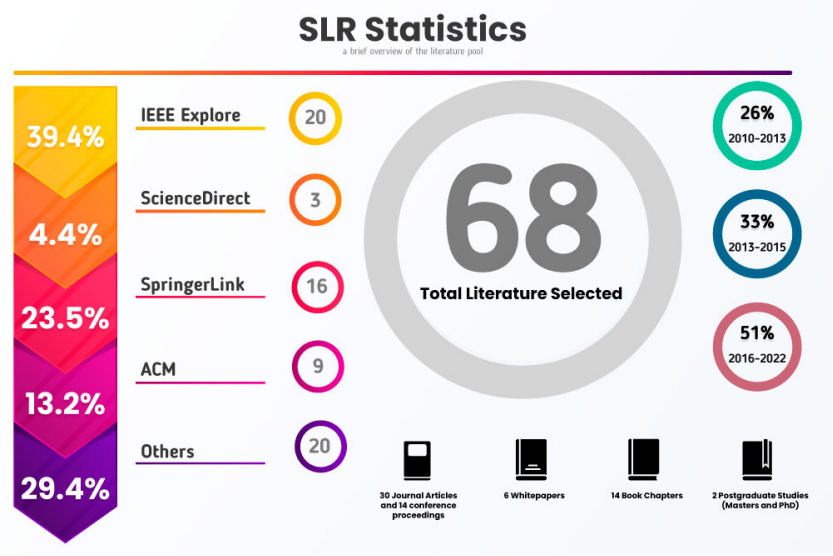

Figure 2: SLR statistics

By this stage, the research objective is set, studies are pooled, assessed and refined, thus the research embarked on the actual synthesis of data. To increase transparency, RAs and standards found as the result of this SRL is presented in table 1. For this purposes, the software Nvivo [57] has been used to code, label, and classify studies. Initially, all the keywords aforementioned has been created as nodes in the software, which are then associated to relevant sentences in studies. After coding all the studies, the findings have been synthesized to create theories, which in turn emerged themes and patterns. The findings gained from this SLR grounded the foundation for various aspect of the SLR development.

Study	Author	Year	Type
Towards a big Data reference architecture	[58]	2013	Master's Dissertation
A reference architecture for Big Data solutions introducing a model to perform predictive analytics using Big Data technology	[59]	2013	Conference Paper
A proposal for a reference architecture for long-term archiving, preservation, and retrieval of Big Data	[60]	2014	Conference Paper
Questioning the Lambda architecture; Kappa Architecture	[39]	2014	Practice
Defining architecture components of the Big Data Ecosystem	[61]	2014	Conference Paper
Big Data driven e-commerce architecture	[62]	2015	Journal Article
The solid architecture for real-time management of big semantic data	[63]	2015	Journal Article
Reference architecture and classification of technologies, products and services for big data systems	[64]	2015	Journal Article
A Reference Architecture for Big Data Systems	[65]	2016	Conference Paper
A reference architecture for Big Data systems in the national security domain	[66]	2016	Conference Paper

	88	6344	

Table 1: RAs and Standards found by the result of the SLR

3.4. Construction of the RA

Based on the themes, theories, and patterns realized in the previous steps, the process of RA construction took place. Integral to this step was the identification of elements that the RA should contain, how these elements should be synthesized, and how the RA can be portrayed and communicated. To describe our RA, we followed ISO/IEC/IEEE 42010 standard [67]. This standard pivots on concrete architectures, so we did not 100% conform to it, but rather the good and relevant parts of it has been taken. For instance, architecture viewpoints, statement of corresponding rules, and expression of the architecture through architecture description languages (ADLs) have had direct aspects on the construction of this RA.

A key challenge in the development of this RA was to strike a balance between the specificity of the micro patterns and approaches to system development and general architectural concepts that reflect a view of a the system as an array of interrelated entities. Angelove et al [68] approached this problem by the means of interrogative through a defined framework that aims to guide the creation of RAs. Cloutier et al [17] suggest that a RA should entail technical, business and customer context views, whereas Vogel et al [69] provided classifies RA views based on the usage context, as industry specific, platform specific, industry crosscutting and product line RAs.

Stricker et al [70] expressed their pattern-based RA by adhering several distinct views into one. Chang et al [71] presented NIST BD RA as system constituent of logical components connected though interoperability interfaces in several fabrics. On the other hand, ISO/IEC/IEEE 42010 refrains from using phrases such as “technical architecture”, “physical architecture”, or “business architecture”.

Taking the best evidence from the available body of knowledge, We decided to adhere several views into one and it express the RA through a multi-layer modeling language called Archimate. Archimate is mature modeling language developed by the Open Group that provides with a uniform representation of high-level architectural diagram aimed at portraying and delineating Enterprise architecture [72]. Archimate being listed as a standard architecture description language in ISO/IEC/IEEE 42010, is designed based on a set of related concepts that are specialized towards the system at different architectural layers. This means that the architect is enhanced with an integrated architectural approach that visualizes and describes different architecture domains and their underlying relations [73]; [74].

Archimate utilizes service-orientation to distinguish and relate the application, business and technology layer and use realization relationships to create relationship between concrete elements and more abstract elements across three layers. In addition, Archimate can be customized to account for varying needs of the architect.

3.5. Enabling RA with variability

Enabling RA with variability is an important process that helps with the instantiation of it. This allows RA to remain useful as a priori artefact when it comes to country-specific regulations, and organizational compliance that constrain the architect design decisions [75].

Variability management has been studied in the domain of Business Process Management (BPM) [76]; ;[77]; [78] and Software Product Line Engineering (SPLE) [79][80]; [81]; [82]; [83]. In BPM, variability management revolves around efficient handling of of different variants in business processes, whereas in SPLE, variability management is about modifying and extending the software artefact to account of the requirements of a specific context.

Clear identification of variability and explicit communication of it improves communication between stakeholders, allows for traceability between variation causes and effects and facilitates the decision making [84].

Variation points are decided based on the data collected in previous steps. Galster et al [24] sugguest that there are three approaches to enabling variability;

1. Annotation of the RA
2. Variability views
3. Variability models

We could not find an in-detail explanation of how one should choose the appropriate variability enabling approach. Therefore, inspired by the works of Rurua et al [75], we decided to extend the RA with variability, by the means of Archimate annotations. We have achieved this in two steps; first we developed a custom layer that represents focal variability concepts, and then we extended the RA through annotation. The aim of this process is not find all variability points that may emerge in the usage context, but to provide with high-level system related architectural variabilities that an architect may consider for improvement of design and adoption of the RA.

The variability model is depicted in Fig 3 by the means of Archimate's motivation layer. This modeling is driven by the works of Pohl et al [79] and in specific their graphical notation of variability information, and Rurua et al [75] and in specific, their variability management concepts model.

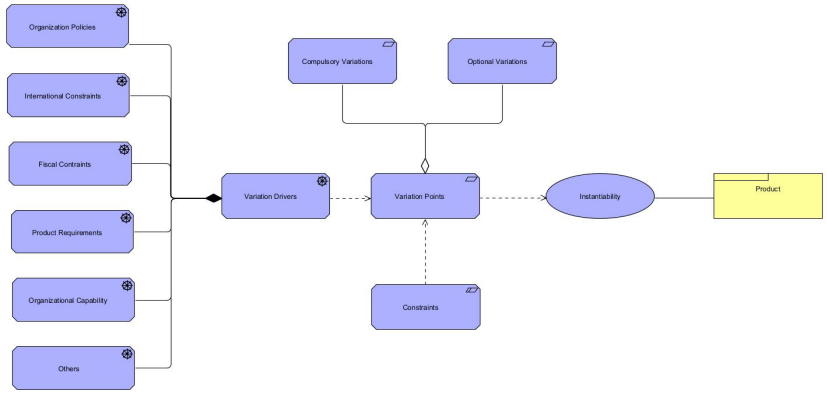

Figure 3: Variability management concepts model

3.6. Evaluation of the RA

Evaluation of the RA is to ensure that it has achieved the goals stated prior to development, to test is effectiveness and usability, and to make sure that it addresses the identified problems. Two fundamental pillars of the evaluation is the correctness and the utility of the RA and how efficient it can be adapted and instantiated [24]. The quality of RA can be assessed by how it can be transformed into an effective organization-specific concrete architecture. The fact that this RA is built upon former RAs helps making the evaluation steps easier as the research can get inspiration from other studies and their approach to evaluation [85].

Nevertheless, evaluation of the RAs is a well-known challenge among researchers [34]; [86]; [87]; [58]. RAs and concrete architectures have distinct qualities. They vary in at least 3 major ways;

1. RAs are of higher level of abstraction
2. in RAs stakeholders are not clearly grouped
3. RAs tend to be focused more on architectural qualities

While there are many well-established methods for assessing concrete architectures such as Scenario-based Architecture Analysis Method [88], Architecture Level Modifiability Analysis [89], Performance Assessment of Software Architecture [90], Architecture Trade-off Analysis Method [91], none of these methods can be directly applied to evaluate RAs. To support this statement, three major issues have been identified. These issues are as follows;

1. One of the main problems for applying existing evaluation methods to RA is the lack of clearly defined group of stakeholders [30], while ATAM and other methods are highly dependent on participation of stakeholders for evaluation. Due to the level of abstractness of the RA, reaching various group of stakeholders and persuade them to anticipate in the study, is problematic and does not fit to the timeline of this study. Even more notably, it is unlikely that all stakeholders will unite around a common

reference architecture as different members may or may not agree with the overall idea of the RAs, may come from different backgrounds, and may lack architectural visions

2. Evaluation frameworks and methods for concrete architectures make use of scenarios. Howbeit due to RAs level of abstraction, creation of usable scenario is difficult. Either a large set of scenarios should be developed covering all the aspects of the RA with regards to specific domain, or a more general scenarios should be developed to cover all the aspects. In the first approach, a large number of scenarios, makes data analysis troublesome and a tedious process. Moreover, the order of prioritization of these scenarios and defining them, and validating them is a problematic task. In the second approach, due to the generality of the scenario, evaluation of effectiveness and usability of the RA becomes difficult and may become incomplete [86]. These challenges have been observed even in the evaluation of highly complex concrete architectures in information systems domain [92].

Based on the problems discussed above, available methods of architecture analysis are not sufficient in evaluating the RA. This has been addressed by various researchers in the industry.

In one study, Angelov et al [30], modified ATAM and extended it to resonate well with RAs. This process took place by invitation of representatives from leading industries for the evaluation process, and the selection of various contexts and defined scenarios for these contexts. Furthermore, ATAM has been extended to evaluate completeness, buildability and applicability. Howbeit the selection of the right candidate and involving them in the process is a time-consuming and daunting task and may yield incomplete information. In addition, candidates maybe lacking architectural visions, increasing the threat to validity.

In addition to extending ATAM for RAs, Graaf et al [93] presented an evaluation approach in which SAAM is extended to help reduce the organizational

impact of it. In Another study by Maier et al, Maier et al. (2013) as a post-graduate thesis in Eindhoven University of Technology, the evaluation of the RA has been conducted by mapping it against existing concrete architectures described in industrial whitepapers and reports. Along the lines, Galstar et al [24] suggested reference implementations, prototyping and incremental approach for the validation of the RA.

Rohling et al [94] have evaluated their RA by mapping it against the requirements set for the study. This was facilitated by the RA research methodology created by Nakagawa et al [95] and the complementary RAModel [26].Inspired by all the studies listed, for the purposes of this study, we will first create a prototype of the RA in an actual organizational setup and then we will use ATAM to evaluate the concrete architecture.

4. Cybermycelium: A Domain Driven Distributed Reference Architecture for Big Data Systems

References

[1] P. Ataei, A. T. Litchfield, Big data reference architectures, a systematic literature review.

[2] B. B. Rad, P. Ataei, The big data ecosystem and its environs, International Journal of Computer Science and Network Security (IJCSNS) 17 (3) (2017) 38.

[3] P. Ataei, A. Litchfield, Neomycelia: A software reference architecturefor big data systems, in: 2021 28th Asia-Pacific Software Engineering Conference (APSEC), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 452–462. doi:10.1109/APSEC53868.2021.00052.
URL https://doi.ieeecomputersociety.org/10.1109/APSEC53868.2021.00052

[4] I. L. Stats, Internet live stats (2019).
URL https://www.internetlivestats.com/

[5] M. Lycett, 'datafication': Making sense of (big) data in a complex world (2013).

[6] B. B. Rada, P. Ataeib, Y. Khakbizc, N. Akbarzadehd, The hype of emerging technologies: Big data as a service.

[7] M. Huberty, Awaiting the second big data revolution: from digital noise to value creation, Journal of Industry, Competition and Trade 15 (1) (2015) 35–47.

[8] M. technology review insights in partnership with Databricks, Building a high-performance data organization (2021).
URL https://databricks.com/p/whitepaper/mit-technology-review-insights-report

[9] N. Partners, Big data and ai executive survey 2021 (2021).
URL https://www.supplychain247.com/paper/big_data_and_ai_executive_survey_2021/pragmadik

[10] M. Analytics, The age of analytics: competing in a data-driven world, Tech. rep., Technical report, San Francisco: McKinsey & Company (2016).

[11] H. Nash, Cio survey 2015, Association with KPMG.

[12] N. Singh, K.-H. Lai, M. Vejvar, T. Cheng, Big data technology: Challenges, prospects and realities, IEEE Engineering Management Review.

[13] B. B. Rad, P. Ataei, Evaluating major issues regarding reliability management for cloud-based applications, IJCSNS 17 (7) (2017) 168.

[14] I. Gorton, J. Klein, Distribution, data, deployment, STC 2015 (2015) 78.

[15] S. Nadal, V. Herrero, O. Romero, A. Abelló, X. Franch, S. Vansummeren, D. Valerio, A software reference architecture for semantic-aware big data systems, Information and software technology 90 (2017) 75–92.

[16] R. K. Len Bass, Dr. Paul Clements, Software Architecture in Practice (SEI Series in Software Engineering) 4th Edition, Addison-Wesley Professional; 4th edition, 2021.

[17] R. Cloutier, G. Muller, D. Verma, R. Nilchiani, E. Hole, M. Bone, The concept of reference architectures, Systems Engineering 13 (1) (2010) 14–27.

[18] J. Kohler, T. Specht, Towards a secure, distributed, and reliable cloud-based reference architecture for big data in smart cities, in: Big Data Analytics for Smart and Connected Cities, IGI Global, 2019, pp. 38–70.

[19] M. Derras, L. Deruelle, J.-M. Douin, N. Levy, F. Losavio, Y. Pollet, V. Reiner, Reference architecture design: A practical approach, in: ICSOFT, pp. 633–640.

[20] J. Bayer, T. Forster, D. Ganesan, J.-F. Girard, I. John, J. Knodel, R. Kolb, D. Muthig, Definition of reference architectures based on existing systems, Fraunhofer IESE, March.

[21] V. Stricker, K. Lauenroth, P. Corte, F. Gittler, S. De Panfilis, K. Pohl, Creating a reference architecture for service-based systems–a pattern-based approach, in: Towards the Future Internet, IOS Press, 2010, pp. 149–160.

[22] E. Gamma, R. Helm, R. Johnson, R. E. Johnson, J. Vlissides, et al., Design patterns: elements of reusable object-oriented software, Pearson Deutschland GmbH, 1995.

[23] E. Y. Nakagawa, R. M. Martins, K. R. Felizardo, J. C. Maldonado, Towards a process to design aspect-oriented reference architectures, in: XXXV Latin American Informatics Conference (CLEI) 2009, 2009.

[24] M. Galster, P. Avgeriou, Empirically-grounded reference architectures: a proposal, in: Proceedings of the joint ACM SIGSOFT conference–QoSA and ACM SIGSOFT symposium–ISARCS on Quality of software

architectures–QoSA and architecting critical systems–ISARCS, 2011, pp. 153–158.

[25] E. Y. Nakagawa, M. Guessi, J. C. Maldonado, D. Feitosa, F. Oquendo, Consolidating a process for the design, representation, and evaluation of reference architectures, in: 2014 IEEE/IFIP Conference on Software Architecture, IEEE, 2014, pp. 143–152.

[26] E. Y. Nakagawa, F. Oquendo, M. Becker, Ramodel: A reference model for reference architectures, in: 2012 Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture, IEEE, 2012, pp. 297–301.

[27] J. F. M. Santos, M. Guessi, M. Galster, D. Feitosa, E. Y. Nakagawa, A checklist for evaluation of reference architectures of embedded systems (s)., in: SEKE, Vol. 13, 2013, pp. 1–4.

[28] M. Derras, L. Deruelle, J. M. Douin, N. Levy, F. Losavio, Y. Pollet, V. Reiner, Reference architecture design: a practical approach, in: 13th International Conference on Software Technologies (ICSOFT), SciTePress-Science and Technology Publications, 2018, pp. 633–640.

[29] I. WG, Iso/iec 26550: 2015-software and systems engineering–reference model for product line engineering and management, ISO/IEC, Tech. Rep.

[30] S. Angelov, J. J. Trienekens, P. Grefen, Towards a method for the evaluation of reference architectures: Experiences from a case, in: European Conference on Software Architecture, Springer, 2008, pp. 225–240.

[31] S. Angelov, J. J. Trienekens, P. Grefen, Extending and adapting the architecture tradeoff analysis method for the evaluation of software reference architectures.

[32] R. Kazman, M. Klein, M. Barbacci, T. Longstaff, H. Lipson, J. Carriere, The architecture tradeoff analysis method, in: Proceedings. fourth ieee

international conference on engineering of complex computer systems (cat. no. 98ex193), IEEE, 1998, pp. 68–78.

[33] S. Angelov, P. Grefen, D. Greefhorst, A classification of software reference architectures: Analyzing their success and effectiveness, in: 2009 Joint Working IEEE/IFIP Conference on Software Architecture & European Conference on Software Architecture, IEEE, 2009, pp. 141–150.

[34] S. Angelov, P. Grefen, An e-contracting reference architecture, Journal of Systems and Software 81 (11) (2008) 1816–1844.

[35] D. Hollingsworth, U. Hampshire, Workflow management coalition: The workflow reference model, Document Number TC00-1003 19 (16) (1995) 224.

[36] H. Zimmermann, Osi reference model-the iso model of architecture for open systems interconnection, IEEE Transactions on communications 28 (4) (1980) 425–432.

[37] OATH, Oath reference architecture, release 2.0 initiative for open authentication, OATH.
URL `https://openauthentication.org/wp-content/uploads/2015/09/ReferenceArchitectureVersion2.pdf`

[38] A. L. Pope, The CORBA reference guide: understanding the common object request broker architecture, Addison-Wesley Longman Publishing Co., Inc., 1998.

[39] J. Kreps, Questioning the lambda architecture, Online article, July 205.

[40] M. Press, L. Joyner, G. Malcolm, Application Architecture for. NET: Designing Applications and Services, Microsoft Press, 2002.

[41] D. Greefhorst, P. Gehner, Achmea streamlines application development and integration, Via Nova Architectura.

[42] D. Greefhorst, Een applicatie-architectuur voor het web bij de bank—de pro's en contra's van toestandsloosheid, Software Release Magazine 2.

[43] H. Wu, A reference architecture for Adaptive Hypermedia Applications, Citeseer, 2002.

[44] A. H. Norta, Exploring dynamic inter-organizational business process collaboration (2007).

[45] Y. Gan, Y. Zhang, D. Cheng, A. Shetty, P. Rathi, N. Katarki, A. Bruno, J. Hu, B. Ritchken, B. Jackson, et al., An open-source benchmark suite for microservices and their hardware-software implications for cloud & edge systems, in: Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, 2019, pp. 3–18.

[46] R. Laigner, Y. Zhou, M. A. V. Salles, Y. Liu, M. Kalinowski, Data management in microservices: State of the practice, challenges, and research directions, arXiv preprint arXiv:2103.00170.

[47] N. Dragoni, S. Giallorenzo, A. L. Lafuente, M. Mazzara, F. Montesi, R. Mustafin, L. Safina, Microservices: yesterday, today, and tomorrow, Present and ulterior software engineering (2017) 195–216.

[48] Qcon software conferences (2022).
URL https://qconferences.com/

[49] State of data mesh 2022 (2022).
URL https://www.thoughtworks.com/about-us/events/state-of-data-mesh-2022

[50] Worldwide software architecture summit'21 (2021).
URL https://events.geekle.us/software_architecture/

[51] Kafka summit europe 2021 (2021).
URL https://www.confluent.io/events/kafka-summit-europe-2021/

[52] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. A. Stewart, Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement, Systematic reviews 4 (1) (2015) 1–9.

[53] C. Castellanos, B. Perez, D. Correal, Smart transportation: A reference architecture for big data analytics, in: Smart Cities: A Data Analytics Perspective, Springer, 2021, pp. 161–179.

[54] G. M. Sang, L. Xu, P. d. Vrieze, Simplifying big data analytics systems with a reference architecture, in: Working Conference on Virtual Enterprises, Springer, 2017, pp. 242–249.

[55] I. O. for Standardization (ISO/IEC), Iso/iec 20546:2019 (2019).
URL https://www.iso.org/standard/68305.html

[56] I. O. for Standardization (ISO/IEC), Iso/iec tr 20547-1:2020 (2020).
URL https://www.iso.org/standard/71275.html

[57] Unlock insights in your data with the best qualitative data analysis software (2022).
URL https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

[58] M. Maier, A. Serebrenik, I. Vanderfeesten, Towards a big data reference architecture, University of Eindhoven.

[59] B. Geerdink, A reference architecture for big data solutions introducing a model to perform predictive analytics using big data technology, in: 8th international conference for internet technology and secured transactions (ICITST-2013), IEEE, 2013, pp. 71–76.

[60] P. Viana, L. Sato, A proposal for a reference architecture for long-term archiving, preservation, and retrieval of big data, in: 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications, IEEE, 2014, pp. 622–629.

[61] Y. Demchenko, C. De Laat, P. Membrey, Defining architecture components of the big data ecosystem, in: 2014 International conference on collaboration technologies and systems (CTS), IEEE, 2014, pp. 104–112.

[62] A. Ghandour, Big data driven e-commerce architecture, International Journal of Economics, Commerce and Management 3 (5) (2015) 940–947.

[63] M. A. Martínez-Prieto, C. E. Cuesta, M. Arias, J. D. Fernández, The solid architecture for real-time management of big semantic data, Future Generation Computer Systems 47 (2015) 62–79.

[64] P. Pääkkönen, D. Pakkala, Reference architecture and classification of technologies, products and services for big data systems, Big data research 2 (4) (2015) 166–186.

[65] G. M. Sang, L. Xu, P. De Vrieze, A reference architecture for big data systems, in: 2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), IEEE, 2016, pp. 370–375.

[66] J. Klein, R. Buglak, D. Blockow, T. Wuttke, B. Cooper, A reference architecture for big data systems in the national security domain, in: 2016 IEEE/ACM 2nd International Workshop on Big Data Software Engineering (BIGDSE), IEEE, 2016, pp. 51–57.

[67] I. International Organization for Standardization (ISO/IEC), Iso/iec/ieee 42010:2011 (2020).
URL `https://www.iso.org/standard/50508.html`

[68] S. Angelov, P. Grefen, D. Greefhorst, A framework for analysis and design of software reference architectures, Information and Software Technology 54 (4) (2012) 417–431.

[69] O. Vogel, I. Arnold, A. Chughtai, E. Ihler, T. Kehrer, U. Mehlig, U. Zdun, Software-architektur: Grundlagen–konzepte, Praxis 2.

[70] V. Stricker, K. Lauenroth, P. Corte, F. Gittler, S. De Panfilis, K. Pohl, Creating a reference architecture for service-based systems-a pattern-based approach, in: Future Internet Assembly, pp. 149–160.

[71] W. L. Chang, D. Boyd, Nist big data interoperability framework: Volume 6, big data reference architecture, Report (2018).

[72] M. Lankhorst, A language for enterprise modelling, in: Enterprise Architecture at Work, Springer, 2013, pp. 75–114.

[73] M. M. Lankhorst, H. A. Proper, H. Jonkers, The anatomy of the archimate language, International Journal of Information System Modeling and Design (IJISMD) 1 (1) (2010) 1–32.

[74] W. Engelsman, D. Quartel, H. Jonkers, M. van Sinderen, Extending enterprise architecture modelling with business goals and requirements, Enterprise information systems 5 (1) (2011) 9–36.

[75] N. Rurua, R. Eshuis, M. Razavian, Representing variability in enterprise architecture, Business & Information Systems Engineering 61 (2) (2019) 215–227.

[76] M. La Rosa, W. M. van der Aalst, M. Dumas, A. H. Ter Hofstede, Questionnaire-based variability modeling for system configuration, Software & Systems Modeling 8 (2) (2009) 251–274.

[77] M. Rosemann, W. M. Van der Aalst, A configurable reference modelling language, Information systems 32 (1) (2007) 1–23.

[78] A. Hallerbach, T. Bauer, M. Reichert, Capturing variability in business process models: the provop approach, Journal of Software Maintenance and Evolution: Research and Practice 22 (6-7) (2010) 519–546.

[79] K. Pohl, G. Böckle, F. Van Der Linden, Software product line engineering: foundations, principles, and techniques, Vol. 1, Springer, 2005.

[80] L. Chen, M. A. Babar, A systematic review of evaluation of variability management approaches in software product lines, Information and Software Technology 53 (4) (2011) 344–362.

[81] K. Schmid, I. John, A customizable approach to full lifecycle variability management, Science of Computer Programming 53 (3) (2004) 259–284.

[82] M. Svahnberg, J. Van Gurp, J. Bosch, A taxonomy of variability realization techniques, Software: Practice and experience 35 (8) (2005) 705–754.

[83] M. Sinnema, S. Deelstra, P. Hoekstra, The covamof derivation process, in: International Conference on Software Reuse, Springer, 2006, pp. 101–114.

[84] K. Czarnecki, P. Grünbacher, R. Rabiser, K. Schmid, A. Wasowski, Cool features and tough decisions: a comparison of variability modeling approaches, in: Proceedings of the sixth international workshop on variability modeling of software-intensive systems, 2012, pp. 173–182.

[85] R. Sharpe, K. Van Lopik, A. Neal, P. Goodall, P. P. Conway, A. A. West, An industrial evaluation of an industry 4.0 reference architecture demonstrating the need for the inclusion of security and human components, Computers in industry 108 (2019) 37–44.

[86] P. Avgeriou, Describing, instantiating and evaluating a reference architecture: A case study, Enterprise Architecture Journal 342 (2003) 1–24.

[87] E. Cioroaica, S. Chren, B. Buhnova, T. Kuhn, D. Dimitrov, Towards creation of a reference architecture for trust-based digital ecosystems, in: Proceedings of the 13th European Conference on Software Architecture-Volume 2, pp. 273–276.

[88] R. Kazman, L. Bass, G. Abowd, M. Webb, Saam: A method for analyzing the properties of software architectures, in: Proceedings of 16th International Conference on Software Engineering, IEEE, 1994, pp. 81–90.

[89] P. Bengtsson, N. Lassing, J. Bosch, H. van Vliet, Architecture-level modifiability analysis (alma), Journal of Systems and Software 69 (1-2) (2004) 129–147.

[90] L. G. Williams, C. U. Smith, Pasasm: a method for the performance assessment of software architectures, in: Proceedings of the 3rd international workshop on Software and performance, pp. 179–189.

[91] R. Kazman, M. Klein, M. Barbacci, T. Longstaff, H. Lipson, J. Carriere, The architecture tradeoff analysis method, in: Proceedings. Fourth IEEE International Conference on Engineering of Complex Computer Systems (Cat. No. 98EX193), IEEE, pp. 68–78.

[92] P. Bengtsson, J. Bosch, Scenario-based software architecture reengineering, in: Proceedings. Fifth International Conference on Software Reuse (Cat. No. 98TB100203), IEEE, 1998, pp. 308–317.

[93] B. Graaf, H. Van Dijk, A. Van Deursen, Evaluating an embedded software reference architecture-industrial experience report, in: Ninth European Conference on Software Maintenance and Reengineering, IEEE, 2005, pp. 354–363.

[94] A. J. Rohling, V. V. G. Neto, M. G. V. Ferreira, W. A. Dos Santos, E. Y. Nakagawa, A reference architecture for satellite control systems, Innovations in Systems and Software Engineering 15 (2) (2019) 139–153.

[95] E. Y. Nakagawa, R. M. Martins, K. R. Felizardo, J. C. Maldonado, Towards a process to design aspect-oriented reference architectures, in: XXXV Latin American Informatics Conference (CLEI) 2009, 2009.